Optimizing the Popularity of Twitter Messages through User Categories

Rupert Lemahieu, Steven Van Canneyt, Cedric De Boom and Bart Dhoedt
Department of Information Technology
Ghent University - iMinds
Ghent, Belgium
Email: {steven.vancanneyt,cedric.deboom,bart.dhoedt}@ugent.be

***Abstract*—In this paper, we investigate how the category of a Twitter user can be used to better predict and optimize the popularity of tweets. The contributions of this paper are threefold. First, we compare the influence of content features on the popularity of tweets for different user categories. Second, we present a regression model to predict the popularity of tweets given the content features as input. To construct this model, we interpolate a generic regression model, which is trained on all data, and a category-specific model, which is only trained on tweets from users of the same category as the user of the given tweet. In this way we can combine the advantage of the robustness of a generic model, with the ability of category-specific models to pick up on category-specific influence of content features. The third contribution is the investigation of the feasibility of boosting the popularity of a tweet by setting up an experiment in which we proactively adapt content features in order to optimize the popularity of tweets. Based on this research, we conclude that the introduction of user categories leads to a more precise analysis and better predictions. In the hands-on experiment, we observed a gain in popularity by proactively adapting content features.**

***Keywords*-Twitter, popularity prediction, optimization, user categories.**

I. Introduction

Twitter is one of the most used and powerful social media. Nowadays, over 500 million people are registered on Twitter, of which 302 million are active users[1]. This makes Twitter a useful medium for companies to inform users about their products and services. Users, however, are often flooded with Twitter messages, with the consequence that many tweets are read by almost nobody. This can lead to a drop in popularity and even sales of the company.

Therefore, it is important to determine the optimal content of Twitter posts so that they are read, retweeted and replied by many users. For instance, the responses to the tweets typically rise when they contain links to the blog of the company [1]. In addition, the subject of the tweet, the used hashtags, the attached media—such as images, vines, URLs,...—will affect the number of responses [2], [3], [4], [5], [6], [7], [8]. We will study the influence of the content features on a tweet's popularity, as these are most easily adaptable in comparison to other kinds of features. The analysis about this influence, and the prediction of popularity using these content features, distinguishes itself from other research by categorizing the Twitter users who post the tweets. We express the popularity in terms of the number of retweets or favorites a tweet receives.

The remainder of this paper is structured as follows. We start with a review of related work in Section II. Next, in Section III, details about the data gathering for analysis and prediction is explained. Section IV describes the analysis of the influence of content features on the popularity of tweets for each considered user category. We build a model to predict the popularity of tweets in Section V. Section VI treats a hands-on experiment in which content features are pro-actively optimized in order to boost popularity. Finally, Section VII concludes this paper.

II. Related Work

It has been found that the popularity of a tweet is influenced by some of the tweet's properties [2]. We distinguish a number of feature families. The first ones are the social factors, e.g. the number of followers the user who post the tweet has or the fact if the user is verified or not [3]. Another kind of features are those related to the user history, e.g. the number of tweets placed by the user that were favorited, retweeted or replied to [4]. Next are the timing features, such as the local time of the tweet or the temporal activity of the community [5]. Finally, content features were considered, which include the sentiment value of the tweet [6], hashtags [7], URLs [1], pictures [8] and user mentions [4]. As the content features can most easily adapted for popularity optimization, we will focus on these features in the rest of this paper.

III. Data Gathering and Preprocessing

We make use of the user categories defined by Socialbakers, a social media analytics company.[2] They consider 8 user categories and select the users with the most followers in each category. The first two columnn of Table I gives an overview of the categories along with its top user, i.e. the user that has the most followers in that category. From each category, we consider the 500 users that have most followers. For each user, we crawl the 1000 most recent tweets—or less, if the user has not posted 1000 tweets yet. This

[1] https://about.twitter.com/nl/company

[2] http://www.socialbakers.com/statistics/twitter/

Table I
OVERVIEW OF THE GATHERED DATA PER USER CATEGORY.

Category	Top user Top	#Users	#Tweets
Brands	SamsungMobile	500	479 296
Celebrities	katyperry	500	476 448
Community	UberFacts	500	447 518
Entertainment	SportsCenter	500	489 337
Media	YouTube	500	493 197
Place	MuseumModernArt	500	488 629
Society	BarackObama	500	481 167
Sport	realmadrid	500	494 194
Total		8000	3 849 786

Table II
OVERVIEW OF THE AVERAGE RETWEETS AND FAVORITES PER USER CATEGORY.

Category	Average retweets	Average favorites
Brands	68	49
Celebrities	2112	2685
Communities	828	777
Entertainment	340	341
Media	98	59
Places	25	12
Society	217	917
Sport	99	69
Total	473	614

gathering process was performed on April 2, 2015 which resulted in 3.8 million tweets. The total number of collected tweets per category are presented in the last column of Table I. The average number of favorites and retweets for each category are shown in Table II.

For each tweet, we retrieve five content features, of which 4 can directly be obtained through the Twitter API. These are the number of user mentions, the number of hashtags, the number of URLs and the number of pictures in the tweet. The fifth feature comprises the sentiment of the tweet, which is expressed with a value between 0 and 1; 0 being neutral and 1 being extremely positive or negative. The sentiment of a tweet is calculated as the average of the sentiment values of its words. These word sentiment values are determined based on the LabMT word list, which contains over 10 000 words that were labeled through Amazon's Mechanical Turk [9].

IV. ANALYSIS OF THE INFLUENCE OF CONTENT FEATURES ON THE POPULARITY OF TWEETS

In this section, we determine the influence of content features on the popularity of tweets, and how this differs between tweets of different user categories. In particular, for each user category, we want to rank the content features based on their predictive importance towards the tweet popularity. This is performed using LASSO with least angle regression [10]. LASSO is a shrinkage and selection method for linear regression. It minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients. By decreasing the bound, features will be removed from the model as their coefficients take zero values. The later the feature is removed from the model, the bigger its predictive importance. The sequence of being removed from the model can thus be used to rank the content features based on their predictive value.

Table III
OVERVIEW OF THE IMPORTANCE OF EACH PREDICTOR FOR THE RETWEETS PER USER CATEGORY USING LASSO.

Category	Pictures	Hashtags	Mentions	URLs	Sentiment
Brands	1	2	5	3	4
Celebrities	4	2	1	3	5
Communities	1	2	5	3	4
Entertainment	2	5	1	3	4
Media	3	5	2	1	4
Places	2	3	5	1	4
Society	5	3	2	1	4
Sport	1	3	4	2	5
General	4	3	2	1	5

Table IV
OVERVIEW OF THE IMPORTANCE OF EACH PREDICTOR FOR THE FAVORITES PER USER CATEGORY USING LASSO.

Category	Pictures	Hashtags	Mentions	URLs	Sentiment
Brands	2	3	1	4	5
Celebrities	4	2	1	3	5
Communities	1	3	2	4	5
Entertainment	4	2	1	3	5
Media	1	4	2	3	5
Places	2	3	1	4	5
Society	5	3	1	2	4
Sport	2	4	1	3	5
General	4	3	1	2	5

The results when retweets or favorites are considered as tweet popularity measure are listed in Table III and Table IV, respectively. Several conclusions can be drawn from these results. First of all, the *sentiment* feature has one of the least predictive value of the considered features, as this is consistently ranked fourth or fifth. There are differences between the favorites and the retweets, for instance the *user mentions* are ranked first or second for the favorites, but obtain for most categories a lower ranking for the retweets. This is because users are more likely to favorite a tweet in which they are mentioned, than to retweet it. We see some different rankings among the user categories, which confirms our presumption that introducing user categories would lead to a more precise analysis. For instance, in case of the retweets, the presence of *pictures* is ranked first for *communities*, whereas the presence of a *URL* is ranked first for the *media* category. This could be due to the fact that most of the considered *communities* post funny facts, jokes and memes. For those posts, having a picture in the tweet has a high impact on its popularity. The *media* mainly contains Twitter accounts of newspapers and other news channels. The most popular tweets that they post contain headlines including a link to their website where the whole article is available to read, leading to a high predictive value of the *URL* feature.

Table V
GENERIC MODEL: OVERVIEW OF CONSIDERED REGRESSION METHODS ALONG WITH THE OBTAINED MSLE ON THE VALIDATION SET.

Method	MSLE (Retweets)
Ridge Regression	18.048
Principal Component Regression	9.844
Partial Least Squares Regression	9.521
Least Squares Regression	9.479
Generalized Additive Model	9.479
Multivariate Adaptive Regression Splines	8.478

Table VI
MSLE VALUES OF THE PREDICTIONS ON THE TEST SET USING THE GENERIC TRAINED MODEL.

	MSLE	
Category	Retweets	Favorites
Brands	15.136	14.207
Celebrities	8.586	12.646
Communities	5.285	8.628
Entertainment	8.578	10.661
Media	12.083	12.340
Places	5.762	7.940
Society	6.175	9.768
Sport	5.975	9.444
Average	8.448	10.704

V. PREDICTING THE POPULARITY OF TWEETS

After analyzing the importance of each content feature to predict the popularity of a tweet, we tackle the problem of predicting the popularity of a tweet, given its features. Therefore, we construct regression models consisting of a generic model and a category-specific model. The models are trained using a training set, containing 50% of the tweets of each user, optimized using the validation set, containing 25% of the data, and evaluated using the test set, containing the remaining data. To express performance, the Mean Squared Logarithmic Error (MSLE) is used, which is defined as

$$\text{MSLE} = \sum_{i=1}^{n} \left(\log\left(1 + g(x_i)\right) - \log\left(1 + \hat{g}(x_i)\right)\right)^2$$

with n the number of tweets in the test set, x_i the observed values for the features of tweet i, $g(x_i)$ the actual number of retweets (resp. favorites) of tweet i, and $\hat{g}(x_i)$ the predicted number of retweets (resp. favorites). Note that we also consider the number of follower as a feature in this section, as this feature is a very good predictor of a tweet's popularity [3]. Our approach is explained in more detail in the rest of this section.

A. Generic Model

The generic model is trained on all data in the training set, without taking into account the user categories. We consider this generic model as the baseline. The validation set is used to test several regression methods, as shown for retweets in Table V. Multivariate adaptive regression splines (MARS) [11] performs best as it has the lowest MSLE of all methods. This regression method is considered in the rest of this paper.

Table VII
MSLE VALUES OF THE PREDICTION ON THE TEST SET USING THE CATEGORY-SPECIFIC MODELS.

	MSLE	
Category	Retweets	Favorites
Brands	6.419	5.168
Celebrities	12.060	14.798
Communities	6.345	7.130
Entertainment	10.624	11.359
Media	4.533	4.770
Places	4.854	2.528
Society	5.462	5.115
Sport	4.597	4.536
Average	6.862	6.925

Table VIII
IMPROVEMENT OF THE MSLE VALUES OF THE CATEGORY-SPECIFIC MODELS VERSUS THE GENERIC MODEL.

	Retweets		Favorites	
Category	Absolute	Relative	Absolute	Relative
Brands	8.718	57.59%	9.039	63.62%
Celebrities	-3.475	-40.47%	-2.153	-17.03%
Communities	-1.060	-20.06%	1.498	17.36%
Entertainment	-2.046	-23.86%	-0.698	-6.54%
Media	7.551	62.49%	7.571	61.35%
Places	0.908	15.76%	5.412	68.16%
Society	0.713	11.54%	4.653	47.63%
Sport	1.378	23.06%	4.909	51.98%
Average	1.586	18.77%	3.779	35.30%

We present the obtained MSLE values for the prediction on the test set in Table VI. Notice that the prediction errors vary highly: the *communities* category only has a MSLE of 5.285 for the retweets, while the *brands* category has a MSLE of 15.136 for the retweets. For all but the *brands* category, the prediction error in number of retweets is smaller than for the number of favorites.

B. Category-specific Model

For each considered user category, we constructed a MARS model which is only trained on tweets which are posted by a user of the category. For each tweet in the test set, we first decide to which category it belongs, and then use the corresponding model to predict the number of retweets (resp. favorites). The resulting MSLE values are shown in Table VII. The easiest way to interpret these results is to compare them to the results of the generic baseline model. The improvement of the predictions on the test set of the category-specific models versus the generic model is expressed in Table VIII.

Comparing the MSLE values with the generic model, the category-specific models perform on average better than the generic model. We see that especially the favorites benefit from the introduction of user categories. In a few cases, the category-specific models does not outperform the generic model. A possible explanation could be that the training set is not big enough—the generic model uses a training set 8 times as large—which can result in overfitting the data.

Table IX
MSLE VALUES OF THE PREDICTION ON THE TEST SET USING THE SMOOTHED MODELS, TOGETHER WITH SMOOTHING PARAMETER α.

	Retweets		Favorites	
Category	α	MSLE	α	MSLE
Brands	0.001	6.417	0.001	5.158
Celebrities	0.969	8.429	0.987	12.567
Communities	0.993	5.265	0.009	6.917
Entertainment	1.000	8.578	0.966	10.156
Media	0.000	4.533	0.000	4.770
Places	0.835	4.634	0.000	2.528
Society	0.000	5.462	0.000	5.115
Sport	0.066	4.415	0.000	4.536
Average		5.967		6.468

Table X
IMPROVEMENT OF THE MSLE VALUES OF THE SMOOTHED/ MODELS VERSUS THE GENERIC MODEL.

	Retweets		Favorites	
Category	Absolute	Relative	Absolute	Relative
Brands	8.719	57.61%	9.049	63.69%
Celebrities	0.156	1.82%	0.078	0.62%
Communities	0.020	0.38%	1.711	19.83%
Entertainment	0.000	0.00%	0.505	4.73%
Media	7.551	62.49%	7.571	61.35%
Places	1.128	19.57%	5.412	68.16%
Society	0.713	11.54%	4.653	47.63%
Sport	1.560	26.10%	4.909	51.98%
Average	2.481	29.37%	4.236	39.57%

C. Smoothed Model

As neither the category-specific models or the generic model perform best in all categories, a logical next step is to create a new model per category that contains best elements of both: the large training set of the generic model and the specificity of the category-specific model. This is realized through smoothing, in which a weighted version of the category-specific model and the generic model is taken into consideration. If we denote $\mathcal{P}$ as a prediction, the final prediction $\mathcal{P}_{final}$ can be expressed as:

$$\mathcal{P}_{final} = \alpha \cdot \mathcal{P}_{generic} + (1 - \alpha) \cdot \mathcal{P}_{category}$$

where $\mathcal{P}_{generic}$ denotes the prediction by the generic model, $\mathcal{P}_{category}$ denotes the prediction by the category-specific model and $\alpha \in [0; 1]$ the smoothing parameter. The smoothing parameter α is optimized per category on the validation set. The resulting mean squared logarithmic error on the test set are shown in Table IX, along with the smoothing parameters used to obtain these values. To easier interpret these values, we compare the performance of the smoothed models with the generic baseline model in Table X.

On average, in comparison with the generic baseline model, the predictions of the smoothed model improve with almost 30% for the retweets and 40% for the favorites. We can distinguish a number of categories which benefit from a large improvement. For instance, an improvement of about 60% is obtained for the *brands* and *media* category, both for number of retweets and favorites. The improvement is limited for categories such as *celebrities*, *communities* and *entertainment*. One of the reasons is that the tweets of these categories receives on average a lot of retweets and favorites, as shown in Table II. For those tweets, it is harder to improve the mean squared logarithmic error. In general, we can conclude based on our experiments that the introduction of user categories lead to better popularity predictions.

Table XI
OVERVIEW OF THE CHOSEN ACCOUNT PER CATEGORY.

Category	User
Brands	Microsoft
Celebrities	katyperry
Community	UberFacts
Entertainment	hootsuite
Media	cnnbrk
Place	WaltDisneyWorld
Society	BarackObama
Sport	nfl

VI. PROACTIVELY OPTIMIZING THE POPULARITY OF TWEETS

As stated above, we focus on content features as these can most easily be adapted to improve the popularity of tweets. To confirm this hypothesis, we investigate to what extent it is possible to boost the popularity of a tweet by proactively changing its content features. For each considered user category, two Twitter accounts are set up. The baseline accounts post tweets with plain text, and the extended accounts post the same text, but with improved content features such as hashtags and pictures. The tweets that are posted, are based on existing Twitter messages that other users already posted. In each category, one user is picked from whom tweets is used in this experiment, these are shown in Table XI.

Firstly, a baseline tweet without any optimized content features is composed from the original tweet. This is done by stripping the tweet of all its entities, such as hashtags, pictures and user mentions. The next step is to construct an optimized version of this baseline tweet. This accomplished by actively optimizing content features, i.e. enriching the baseline tweet with media and entities. For instance, hashtags and pictures are added to enlarge the probability of our tweet being noticed. Both the baseline and the optimized tweet are tweeted by the two different accounts. We illustrate this approach with an example. Figure 1 depicts an original tweet by *cnnbrk*, Figure 2 shows the baseline tweet based on this tweet and Figure 3 displays the extended tweet.

This experiment was performed between February 2 and March 7, 2015. The total response each category received on the baseline and extended account is listed in respectively Table XII and Table XIII. From these tables, it can be seen that the extended accounts generally receive more interaction, although the differences are small. The advantage of our approach is that the outcome is comparable, since all

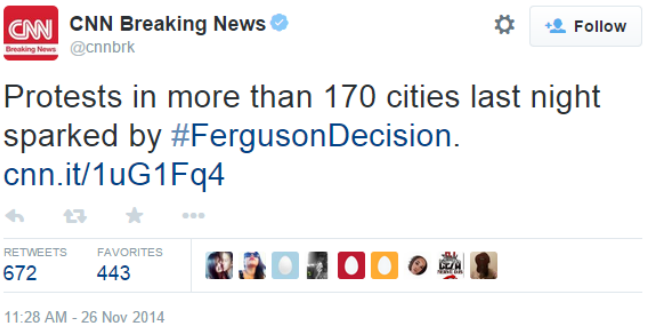

Figure 1. An original tweet from *cnnbrk*.

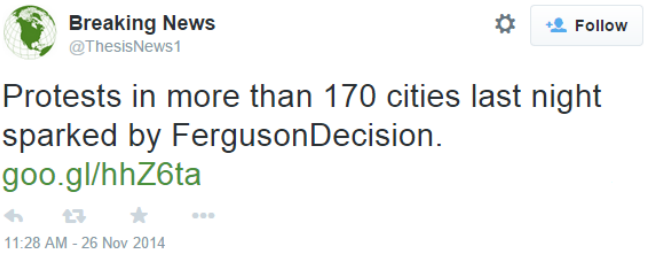

Figure 2. The baseline tweet based on the tweet depicted in Figure 1.

Figure 3. The extended tweet based on the tweet depicted in Figure 1.

accounts start without any followers and post the same, slightly adjusted, tweets. A disadvantage is that because the accounts have no followers, the overall response is small. It might be interesting for the future to conduct this experiment within an actual existing and popular accounts, such as *cnnbrk*. This would probably result in more interaction and gives us a more realistic view of how content features can be adapted to improve the popularity of tweets. Naturally, this would mean that the account needs to agree with the experiment.

Table XII
STATISTICS FOR THE BASELINE ACCOUNTS.

Category	Tweets	Followers	Favorites	Retweets	Replies
Brands	96	0	1	0	0
Celebrities	22	0	0	0	0
Communities	2028	0	8	2	3
Entertainment	360	1	6	2	2
Media	533	3	4	4	1
Places	79	1	2	1	1
Society	61	0	1	0	0
Sport	211	1	19	6	3

Table XIII
STATISTICS FOR THE EXTENDED ACCOUNTS.

Category	Tweets	Followers	Favorites	Retweets	Replies
Brands	96	3	11	10	2
Celebrities	22	0	0	0	0
Communities	2028	1	11	2	4
Entertainment	360	9	36	13	4
Media	533	2	5	7	3
Places	79	2	6	1	0
Society	61	0	1	0	0
Sport	211	2	7	2	0

VII. CONCLUSION

We investigated the impact of considering the user category of a tweet to better predict and optimize its popularity. We focused on content features of the tweet, as these are most easily adaptable to improve the popularity of a tweet. For this analysis, we selected users from 8 categories such as brands, celebrities and media. The tweets of these users were first used to compare the impact of content features on the popularity of tweets for the different user categories. Using a method based on LASSO, we concluded that the most important content features differ for different user categories. Second, we constructed a regression method to demonstrate how user categories can be used to improve the prediction of tweet's popularity. Finally, we have conducted an on-hands experiment on Twitter in which we proactively changed content features of a tweet in order to optimize it's popularity. The result of this experiment seems encouraging at this step. Based on this research, we can conclude that making use of user categories is advantageous for analyzing, predicting and optimizing the popularity of tweets. In future work, we will construct a method which first automatically estimates the category of a given user. This will result in a dataset with much more categories and users, and a method that can be used for all Twitter users, even when their categories are unknown.

ACKNOWLEDGMENT

Steven Van Canneyt is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT). Cedric De Boom is funded by a Ph.D. grant of Ghent University, Special Research Fund (BOF).

REFERENCES

[1] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[2] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *Proceedings of the 2nd International Conference on Social Computing*, 2010, pp. 177–184.

[3] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! Predicting message propagation in Twitter." in *Proceedings of the 5th International Conference on Weblogs and Social Media*, 2011, pp. 586–589.

[4] Y. Artzi, P. Pantel, and M. Gamon, "Predicting responses to microblog posts," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 602–606.

[5] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 1633–1636.

[6] H. Lakkaraju and J. Ajmera, "Attention prediction on social media brand pages," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*.

[7] Z. Ma, A. Sun, and G. Cong, "Will this #hashtag be popular tomorrow?" in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 1173–1174.

[8] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in Twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web*, 2011, pp. 57–58.

[9] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PloS one*, vol. 6, no. 12, pp. 1–26, 2011.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[11] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, pp. 1–67, 1991.